https://doi.org/10.1093/nargab/lqac059

A bioinformatic-assisted workflow for genome-wide identification of ncRNAs

Matthias Schmal[1], Crystal Girod[2], Debbie Yaver[2], Robert L. Mach[3] and Astrid R. Mach-Aigner[1,3,*]

[1]Christian Doppler laboratory for optimized expression of carbohydrate-active enzymes, Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorfer Str. 1A, Vienna A-1060, Austria, [2]Production Strain Technology, Novozymes Inc., California, Davis, USA and [3]Institute of Chemical, Environmental and Bioscience Engineering, TU Wien, Gumpendorfer Str. 1A, Vienna A-1060, Austria

Received March 02, 2022; Revised June 29, 2022; Editorial Decision July 25, 2022; Accepted August 14, 2022

ABSTRACT

With the upcoming of affordable Next-Generation Sequencing technologies, the number of known non-protein coding RNAs increased drastically in recent years. Different types of non-coding RNAs (ncRNAs) emerged as key players in the regulation of gene expression on the RNA–RNA, RNA–DNA as well as RNA–protein level, ranging from involvement in chromatin remodeling and transcription regulation to post-transcriptional modifications. Prediction of ncRNAs involves the use of several bioinformatics tools and can be a daunting task for researchers. This led to the development of analysis pipelines such as UClncR and lncpipe. However, these pipelines are limited to datasets from human, mouse, zebrafish or fruit fly and are not able to analyze RNA sequencing data from other organisms. In this study, we developed the analysis pipeline Pinc (Pipeline for prediction of ncRNA) as an enhanced tool to predict ncRNAs based on sequencing data by removing transcripts that show protein-coding potential. Additionally, a feature for differential expression analysis of annotated genes as well as for identification of novel ncRNAs is implemented. Pinc uses Nextflow as a framework and is built with robust and well-established analysis tools. This will allow researchers to utilize sequencing data from every organism in order to reliably identify ncRNAs.

INTRODUCTION

The introduction of RNA sequencing changed the view on the complexity of the eukaryotic transcriptome drastically. In particular, the view on the large parts of the genome that do not code for proteins did change from considering this as junk DNA to sequences that presumably fulfill other roles. However, in the early years of RNA biology, the identification and characterization of non-coding RNAs (ncRNAs) was a challenging task.

The group of ncRNAs is very heterogeneous and therefore divided into different classes. Many of those classes are conserved across all domains of life and are known today to play essential roles in the complex cellular machinery, such as RNA splicing, modification of other RNAs, DNA replication (1), dosage compensation, regulation of gene expression in numerous ways (2) and most prominently translation of mRNAs into proteins. All ncRNAs, which do not belong to a specific class and are longer than 200 nucleotides (nt), are termed long non-coding RNAs (lncRNAs). This loose definition leads to lncRNAs being such a heterogeneous group.

With the rapid progression of sequencing technologies in the last two decades, RNAs and especially ncRNAs moved into the focus of the scientific community. Deep sequencing of the whole transcriptome became an affordable tool for many research groups to study gene expression. However, the necessary bioinformatic analysis of large amounts of sequencing data may seem overwhelming. To help researchers with this daunting task, several data analysis pipelines have been developed in recent years such as UClncR (3) and LncPipe (4). These pipelines comprise all necessary data processing and analysis steps to predict lncRNAs from sequencing data. A critical step within this workflow is the distinction between potential coding and non-coding RNAs. By using linguistic features of RNA sequences, a supervised machine learning model can be trained. Since this approach requires curated training data, these models are either trained on intensively studied organisms like human, zebrafish, mouse and fruitfly or using sequence data from

[*]To whom correspondence should be addressed. Tel: +43 1 58801 166558; Fax: +43 1 58801 17299; Email: astrid.mach-aigner@tuwien.ac.at
Present addresses:
Crystal Girod, Beckman Coulter Life Sciences, San Jose, CA 95134, USA.
Debbie Yaver, Nature's Fynd, Chicago, IL, USA.

organisms across all domains of life to obtain a general model. For example, analyzing RNA sequencing data from the industrially applied ascomycete *Trichoderma reesei* was not possible since predictions based on existing programs were not sufficiently precise. For accurate predictions based on RNA sequencing data of this or any other organism a new workflow needed to be established. For the purpose of identification of potentially novel ncRNAs from any organism Pinc was developed. It automates the whole process of analyzing RNA sequencing data to distinguish ncRNAs from coding RNAs.

MATERIALS AND METHODS

Pinc uses state-of-the-art tools to be as robust and versatile as possible.

Pre-processing

Raw sequencing reads need to pass several quality control and processing steps. These steps include adapter removal, read filtering and read trimming. If the output of the pre-processing is of poor quality, results may be compromised. This makes this arguably the most important step in the analysis pipeline. In order to simplify the procedure, Pinc uses fastp, a fast and all-in-one solution that can be customized to meet all possible needs (5).

Transcript assembly

Two ways can be chosen to build transcripts from RNA sequencing data, a *de novo* assembly or a genome-guided approach. As *de novo* assemblies are very complex and often require manual curation of data, they are not suitable for an automated pipeline. Therefore, the reads first need to be aligned against a reference genome. HISAT2, the successor of TopHat, was chosen as alignment tool because it is fast, memory-efficient and sensitive (6). Speed and low memory usage are important considerations especially for RNA sequencing data, which usually consists of large datasets that may even be generated in replicates and/or for different conditions for differential gene expression analysis. After the reads are aligned, they need to be assembled into partial transcripts, also called transfrags. Since Pinc was expected to be suitable for prokaryotic as well as eukaryotic organisms, the assembler needs to accommodate for splice variants of eukaryotic genes. Another consideration is the possible presence of long reads generated by Third-Generation-Sequencers such as Oxford Nanopore or Pacific Biosciences SMRT sequencing technologies. StringTie was incorporated into the pipeline as it can utilize short reads, long reads or a combination of both (7). Using StringTie's -merge mode, the assembled transfrags across of samples and replicates are merged into one non-redundant set to facilitate downstream differential gene expression analysis. As of today, Pinc only supports short reads, however, the option to process long reads can be added at any point without the need to reconstruct major parts of the pipeline.

Filtering of transcripts

Since the goal of Pinc is to predict potentially novel ncRNAs, the constructed transcriptome needs to be filtered. By comparing the assembled transfrags against the provided reference annotation gffcompare assigns each transfrag a tag based on the relationship to each annotated gene (8). Keeping only transfrags, which are labeled with ‘u’ (unknown), ‘i’ (fully contained within an intron) or ‘x’ (exonic overlap on complementary strand to annotated feature), effectively reduces the transcriptome to only transfrags that are either not annotated or have a high chance of being ncRNAs (see gffcompare manual for a detailed description of labels). The remaining transfrags might still contain RNAs coding for a protein, thus, an additional filtering step is applied. However, approaches that include aligning sequences to databases are not feasible for large datasets. Therefore, in recent years many tools expanded on linguistic features of sequences. This allows faster processing of sequences and can even yield better results, especially for lncRNAs. As this is a binary classification problem, there are several machine learning approaches to choose from. The two most straightforward ones would be either a Support-Vector-Machine (SVM), which predicts labels for each RNA, or a regression model, which calculates the probability to be a protein-coding RNA. Due to the extensive research on RNAs, a lot of data for ncRNAs are available to be applied in supervised learning in order to train the model. However, based on the training data, the model might be suitable to predict ncRNAs from humans but might yield rather poor results on other organisms. This led to the availability of models that are trained on well-studied species like *Homo sapiens*, *Drosophila melanogaster* or *Arabidopsis thaliana* on one hand, or models that can predict ncRNAs across all species with the trade-off of slightly lower accuracy. In order to tackle this problem, a novel strategy to reliably predict ncRNAs using data from not exhaustively studied organisms was developed. For this purpose we use a combination of CPC2 (9) and CPAT (10) in order to remove transcripts with protein-coding potential. CPAT uses four linguistic features of known coding and non-coding RNAs and transforms them into scores to build a ‘logistic regression’ model. CPAT comes with four pre-trained models, which are trained on datasets of human, mouse, zebrafish and fruitfly, respectively. In an optimal case, sequence data for mRNAs and ncRNAs of the investigated organism itself is available and can be used to train an organism-specific model. However, in most cases there is not enough information about lncRNAs to train a model specific for a target organism. CPC2 is a general classifier and is used to screen all transcripts for their protein-coding potential. Based on this pre-classification an organism-specific CPAT model is trained. As CPAT applies a logistic regression model to calculate coding potential, a probability cut-off needs to be found to distinguish predicted non-coding from protein-coding RNAs. In the end, all transcripts that show protein-coding potential are removed and in an ideal case only ncRNAs are remaining. During the training of the model, training data undergo 10-fold stratified cross-validation where in every iteration the optimal probability cut-off is calculated using a weighted variant of the Youden's index (11) based on a receiver operating characteristic (ROC) curve. From these 10 iterations, the median of all cut-offs will be calculated and used for the final performance evaluation. The weight can be chosen between 0 and 1. A weight of 0.5 means equal

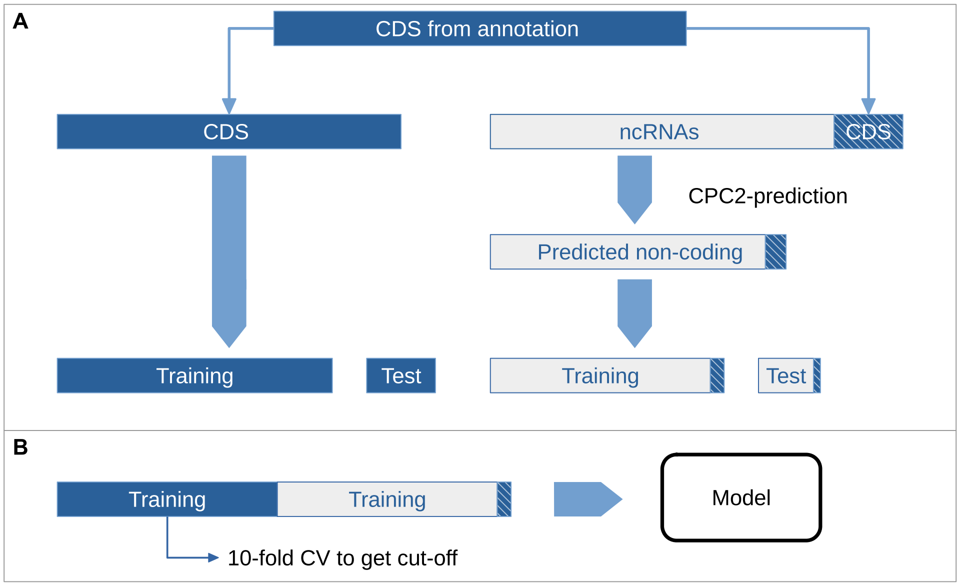

Figure 1. Overview on the process of generation of training data for the test runs. (**A**) Coding sequences (CDS; blue boxes) of *H. sapiens*, *A. thaliana* and *S. cerevisiae* were taken directly from RefSeq. Sufficient data for ncRNAs are available for human and thale cress; whereas the dataset of S. cerevisiae is too small to train CPAT solely on. Therefore, all RefSeq entries of ncRNAs of the phylum ascomycota were used as non-coding training data for CPAT. CDS were split. Right part: a portion of the CDS was used to contaminate the set of ncRNAs (white box) with coding RNAs in a ratio of 5:1 (blue, hatched box). This was done to simulate not annotated coding transcripts, which might still be in the dataset after filtering out all annotated CDS of the genome. CPC2 was used to predict ncRNAs within this mixed RNA pool. 80% of the sequences predicted as non-coding were used as non-coding training set (white box) for CPAT and 20% as the non-coding test set. Left part: Remaining CDS were split again: 80% for the coding training set, 20% for the coding test set. (**B**) The training datasets of ncRNAs (white and hatched boxes) CDS (blue box) were combined in a ratio of 1:1. 10-fold stratified cross-validation (CV) was used to calculate the model-specific 'optimal' cut-off. In each iteration the weighted Youden's index was used to calculate the cut-off. The mean of cut-offs from all 10 iterations is used to predict ncRNAs based on their coding probability calculated by CPAT.

contributions of specificity and sensitivity to determine the cut-off. Decreasing the weight towards 0 will set the cut-off very loose and will lead to higher true-positive rate, however, the false-positive rate will also increase as a trade-off. An increase of the weight will make the cut-off more stringent and predicted lncRNAs will have higher confidence at the cost of potentially missing lncRNAs. To visualize the training process the ROC curves, as well as the performance in dependence on the cut-off, are plotted during each iteration.

Training of classifier

The training of the classifier for Pinc is a two-step process as it combines the potential of both CPC2 and CPAT. After filtering of novel transcripts using gffcompare, the coding potential will be assessed using CPC2. All transcripts, which are predicted as 'non-coding', comprise the non-coding training set for CPAT. To complete the training set, protein-coding transcripts originating from the genome annotation are used. As CPC2 will most likely produce false-positives among those predicted non-coding transcripts, we looked into how these falsely classified transcripts will affect the predictive power of the resulting CPAT classifier. In order to estimate the impact of false-positives of CPC2 in the downstream process, the training set of known ncRNAs was 'contaminated' with known protein-coding sequences. This dataset for ncRNAs contained 80% ncRNAs and 20% protein-coding RNAs. The sequences used to simulate false-positives are removed from the set of CDS originating from the genome annotation. On the 'contaminated' dataset CPC2 was used to predict the coding-potential. Sequences predicted as 'non-coding' will be used to train CPAT. Transcripts predicted as 'coding' by CPC2 will be discarded for the training process only, however, they will be included in the final prediction of the coding potential of novel lncRNAs. A graphical overview of this procedure can be seen in Figure 1A.

In order to evaluate the efficacy of the training, datasets of *H. sapiens*, *A. thaliana* and *ascomycota* were chosen. Since generally little is known about ncRNAs, high-quality datasets are only available for such well-studied organisms. For this reason, the datasets of human and *A. thaliana* comprise exclusively organism-specific mRNAs and ncRNAs. The dataset for ascomycetes is composed of mRNAs from *Saccharomyces cerevisiae* and ncRNAs from all ascomycetes. All datasets are based on RefSeq entries for the respective organism or phylum. The dataset of *ascomycota* is designed to simulate an RNA sequencing experiment to identify potentially novel ncRNAs in a rather unexplored organism. As in most cases, a reference genome of the studied organism is available, while data on ncRNAs are lacking, the combination of mRNAs (or mRNAs based on genome annotation predictions) from the studied organism and ncRNAs from related organisms shall simulate the aforementioned case.

The resulting datasets are split into a training and test set. This allows testing of the trained classifier on data, which it has never 'seen' before. As CPAT only calculates coding probabilities, a cut-off needs to be determined to distinguish

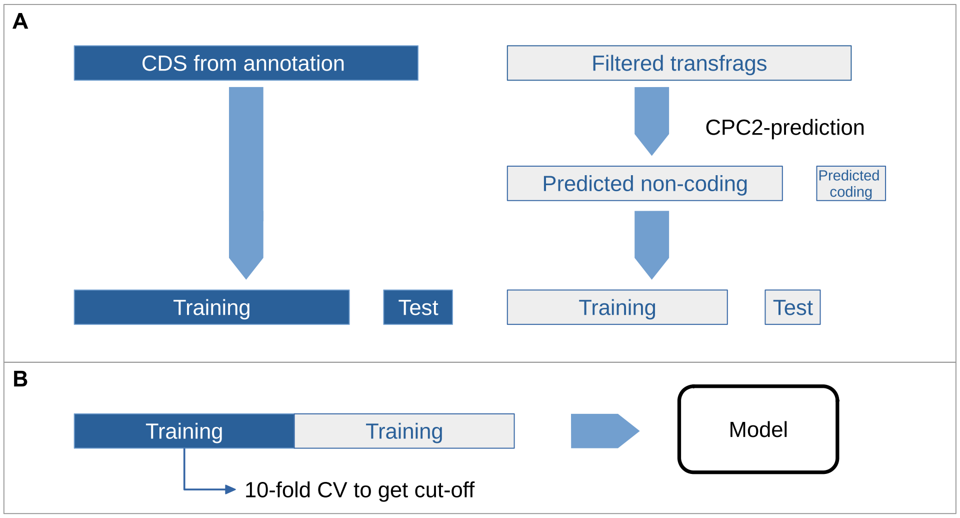

Figure 2. Overview of the process of generation of training data used in Pinc. (**A**) Right part: The filtered transfrags (white box) were subjected to CPC2 prediction of ncRNAs. 80% of the sequences predicted as non-coding were used as non-coding training set for CPAT and 20% as the non-coding test set. Left part: Coding sequences (CDS; blue boxes) are taken from the provided genome annotation and were split: 80% for the coding training set, 20% for the coding test set. (**B**) The training set that consists of ncRNAs (white box) was combined with the CDS training set (blue box) in a ratio of 1:1. 10-fold stratified cross-validation (CV) is used to calculate the model-specific, 'optimal' cut-off. In each iteration the weighted Youden's index was used to calculate the cut-off. The mean of cut-offs from all 10 iterations is used to predict ncRNAs based on their coding probability calculated by CPAT.

coding from non-coding RNAs. In order to estimate such a cut-off, we employ 10-fold stratified cross-validation. The training set will be randomly split again into a training and test subset in such a way, that each resulting subset contains the same ratio of protein-coding to non-coding RNAs. Afterwards, CPAT will be trained on one subset, and the optimal cut-off is calculated using a weighted Youden's Index. This procedure was repeated 10 times. The final cut-off is the median of all tens cut-offs calculated during the cross-validation (see Figure 1B). After training a separate CPAT classifier with each dataset, the three resulting models will be benchmarked against each other, CPC2 and a pretrained CPAT classifier, which was trained on human data.

Pinc employs a similar approach to train the classifier (see Figure 2). Based on the provided genome annotation, Pinc extracts mRNAs and uses them as coding transcripts during the training. For ncRNAs, first a non-redundant transcriptome is built based on all RNA sequencing samples. Using gffcompare all annotated features are removed and the coding probability of the remaining transfrags is assessed using CPC2. Transfrags labelled as 'non-coding' will contain primarily ncRNAs and in most cases also some coding RNAs, which are protein-coding genes missing in the annotation. Based on these datasets CPAT is trained and used to predict the coding potential of all novel transfrags.

Differential expression analysis

Pinc gives the possibility to perform a differential expression analysis using edgeR (12). A basic workflow that allows to compare two conditions is integrated into Pinc. This does not only identify differentially expressed genes, but also differentially transcribed ncRNAs. If replicates for each condition are provided, not only the fold changes of transcripts are reported, but also a statistical significance analysis can be performed. As recommended by the manual, the raw read counts of transcripts should be given to edgeR. For this, the tool HTSeq-count of the HTSeq package is used (13).

Implementation of Pinc

Analysis of sequencing data requires many tools to be installed and maintained. Most tools can only be used through a command-line interface. This can be the first obstacle, since working with those programs can be quite a challenge for users who are not familiar with this environment. Pipelines help researchers by automating running all programs and handle the data flow between those. Pinc is embedded in a Nextflow framework which allows building very intuitive pipelines. However, this still requires that all used programs are installed on the user's system. Therefore, Pinc will be distributed by using a Docker image. This image is a whole-in-one solution as it comes with its own operating system and all required programs installed within the virtual environment. A graphical overview on the whole workflow can be inferred from Figure 3. This allows researchers to use Pinc independent of the operating system and without the tedious installation and setup of multiple programs. In addition, the stand-alone Nextflow pipeline will also be available at GitHub when all necessary programs are already installed on the system.

Generation of test data

RNA sample preparation. During the extensive, large-scale cultivation *Trichoderma reesei* industry strains, such as Iogen-M10, can gradually lose their ability to produce the target product (i.e. cellulases) and transform into a non-producing (cel–) population in contrast to the initial, highly

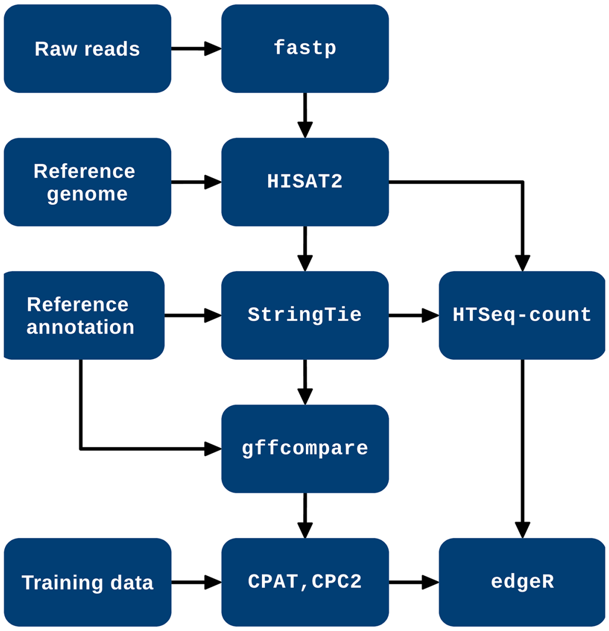

Figure 3. Graphical overview on Pinc. Raw Sequencing reads are filtered based on quality and length using fastp. Subsequently, HISAT2 aligns the reads against the reference genome. StringTie assembles aligned reads into transfrags. Transfrags of already annotated features are removed by filtering for putative novel ncRNAs based on gffcompare's transfrag classification code. Together with the protein-coding RNAs from the reference annotation an organism-specific model is trained using CPC2 and CPAT to assess the coding probability of all putative, novel, non-coding transfrags. As edgeR requires the total count of reads mapped to each transfrag for a differential expression analysis, HTSeq-count was used to count the reads.

productive (cel+) population (14). A (cel+) and a (cel–) strain were cultivated in triplicates in Mandels-Andreotti medium supplemented with 1% (w/v) glucose as the sole carbon source and 0.1% (w/v) peptone or 1% (w/v) lactose and 0.1% (w/v) peptone. These conditions represent low cellulase producing and high cellulase producing conditions, respectively (15). After 48 h of cultivation, mycelium was harvested and total RNA was extracted. To enrich RNAs of interest, an rRNA depletion was performed using 'NEBNext® RNA Depletion Core Reagent Set'. All samples were split into three sets, each set containing a sample from each condition, to ensure fast handling. 50 DNA probes were used to degrade rRNA using RNase H. Probes were designed to leave rRNA fragments of 50–70 nt length after RNase H-mediated degradation. Parameters for purification of samples by size exclusion were chosen as recommended by the manufacturer to obtain RNA fragments of ∼200 nt length.

Library preparation. Library preparation of rRNA depleted samples was done using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina®. RNA was fragmented by incubating at 94°C for 7 min. For PCR enrichment of adapter-ligated DNA 10 cycles was chosen as 500 ng of rRNA-depleted RNA was used in the beginning. Otherwise, the recommended protocol of the manufacturer was followed.

Table 1. Accuracy of differently trained models. This table shows the percentage of correct predictions of each trained model on each test dataset. Rows ('Trained') indicate the organism/phylum of which data were used to train the model. Columns ('Tested') represent the organism/phylum of which the dataset was used for the evaluation of the performance of the model. '*Ascomycota*', '*Arabidopsis*', and 'human' refer to the datasets mentioned in the manuscript section Training of classifier. 'CPAT-human' indicates the model, which is pretrained on human data. 'Human test' indicates the CPAT model trained on published human RNA sequencing data as described in the manuscript section Evaluation of trained classifier. '*T. reesei*' indicates the model, which is trained using the RNA sequencing data from *T. reesei* strain Iogen-M10, produced by Pinc. The '*Ascomycota*' model performs overall best. The models 'Human' and 'Human test' perform very similarly, and outperform 'CPAT-human' as well as 'CPC2'. The accuracy of the '*T. reesei*' model is very poor and the model is not suitable for predicting ncRNAs. All models perform worst on the *Arabidopsis* test data in terms of accuracy

Accuracy		Tested			
		Ascomycota	*Arabidopsis*	Human	Mean
Trained	*Ascomycota*	0.9928	0.9894	0.9981	0.9934
	Arabidopsis	0.9458	0.9569	0.9840	0.9622
	Human	0.9624	0.9693	0.9949	0.9755
	CPC2	0.9472	0.9528	0.9681	0.9561
	CPAT-human	0.9241	0.9337	0.9564	0.9380
	Human test	0.9647	0.9734	0.9931	0.9771
	T. reesei	0.5563	0.7629	0.8850	0.7347

Sequencing. Sequencing was done on an Illumina NextSeq 500 system to acquire a depth of about 20 million paired-end reads per sample.

RESULTS AND DISCUSSION

Evaluation of trained classifier

Especially for lncRNAs the line between coding and noncoding linguistic features blurs, which poses a hard challenge for every prediction tool. The first step was to evaluate the impact of false-positive prediction of ncRNAs by CPC2 as well as possible contamination by not annotated protein-coding sequences. We determined the false-positive rate of CPC2 on the protein-coding sequences used to artificially contaminate the training set. In total, the false-positive rate of CPC2 was 2.2%, 1.3% and 9.3% for the human, *A. thaliana* and *ascomycota* datasets, respectively. Those missclassified sequences were kept in the workflow to see how they will affect the final performance. Anyhow, this result demonstrates that even general models have difficulties to reliably predict the coding potential when working with non-model organisms.

In order to evaluate the performence we calculated three different measures, i.e. accuracy, F1-score and informedness. Our analysis shows that an initial filtering step of ncRNAs will result in accuracies of at least 0.945 across all datasets. We compared our three trained models against the results of CPC2 and CPAT-human, a model specifically trained on human data. In Table 1 accuracies of each classifier are given and Table 2 provides the F_1-scores. The table for informedness is provided in the supporting material (Supplementary Table S1).

Across all test datasets, the models trained with prefiltered data outperform CPC2 and CPAT in every perfor-

Table 2. F_1 score of differently trained models. The F1 score is a measure of the accuracy of a given classifier. The best possible score is 1; the worst is 0. The score is impacted by a low true-positive rate and/or a high false-positive rate. True-negative as well as false-negative rates are not taken into account. Rows ('Trained') indicate the organism/phylum of which data were used to train the model. Columns ('Tested') represent the organism/phylum of which the dataset was used for the evaluation of the performance of the model. '*Ascomycota*', '*Arabidopsis*', and 'human' refer to the datasets mentioned in the manuscript section Training of classifier. CPAT-human indicates the model, which is pretrained on human data. 'Human test' indicates the CPAT model trained on published human RNA sequencing data as described in the manuscript section Evaluation of trained classifier. '*T. reesei*' indicates the model, which is trained using the RNA sequencing data from *T. reesei* strain Iogen-M10, produced by Pinc. Also using the F1 score the '*Ascomycota*' model performs best followed by 'Human' and 'Human test'

*F*1 score		Tested			
		Ascomycota	*Arabidopsis*	Human	Mean
Trained	*Ascomycota*	0.9958	0.9941	0.9988	0.9962
	Arabidopsis	0.9678	0.9753	0.9896	0.9776
	Human	0.9779	0.9825	0.9967	0.9857
	CPC2	0.9686	0.9729	0.9795	0.9737
	CPAT-human	0.9546	0.9620	0.9723	0.9630
	Human test	0.9794	0.9850	0.9956	0.9867
	T. reesei	0.6730	0.8510	0.9236	0.8159

mance measure, most notably the models trained on the ascomycete and human dataset (Tables 1 and 2). Even the *A. thaliana* model, which performs worst out of the three, is still better than the control classifiers.

As these tests were run on highly curated data, good results were expected. In order to simulate results that reflect the usage of less curated data, we run Pinc on rRNA-depleted human RNA sequencing samples (16) to assess the predictive power of the resulting CPAT model. Two conditions were chosen, each having three RNA sequencing runs, respectively. The same method of generating training data for the model was used as shown in Figure 1. Two conditions with three RNA sequencing samples respectively were chosen as testing data. Then, the model was tested on all mRNAs and ncRNAs of *H. sapiens*, *A. thaliana* and the *ascomycota* data. The trained model performed very similarly compared to the models trained on the curated data, also outperforming CPC2 and CPAT human. This demonstrates that this method of training results in a model capable of distinguishing between ncRNAs and coding transcripts without the necessity of the availability of highly curated data. Using Pinc we predicted 3902 novel ncRNAs, all longer than 200 nucleotides and therefore classified as lncRNAs. In contrast, we used lncpipe as comparison, which predicted 53 784 lncRNAs based on the same already published human dataset (16) using standard parameters. In the authors' opinion it is not plausible that next to the already identified ncRNAs, there are still >50 000 novel lncRNAs present. We would rather assume that lncpipe yields a high false-positive rate.

LncRNAs, which show protein-coding features such as long open-reading-frames, seem to diminish the predictive power, as the model tries to fit them within the group of ncRNAs. This leads to a blurry separation between coding and non-coding transcripts and thus, likely miss-annotated protein coding gene fragments. Removing those ambiguous transcripts seems to help to correctly predict RNAs that don't show distinctive features of either class.

Results of Pinc using data of *T. reesei*

Before the sequencing reads were fed into Pinc, we assessed the efficiency of the rRNA depletion. Sequencing reads were mapped against the sequences of the 25S, 18S and 5S rRNA retrieved from the RFAM database (17, 18). This reveals that rRNA depletion worked very well on all samples in set 1 with <3% reads aligning to rRNA sequences. However, in samples from sets 2 and 3 there are between 70% and 80% reads belonging to rRNA. Samples from set 2 have the highest fraction of reads aligning to rRNA genes (see Supplementary Table S2). After transfrags belonging to annotated sequences were removed, 3178 sequences remained as putative novel transfrags. The assessment of the coding potential of these putative novel transfrags resulted in 2064 predicted ncRNAs. Additionally, the predictive power of the trained model was tested on the training datasets of *H. sapiens*, *A. thaliana* and the *ascomycota*. The results showed that the model has an average accuracy of 0.7347 over all tested training sets, indicating that the generated model is not suitable to predict ncRNAs efficiently. In cases like this, when the trained classifier fails to convince, Pinc also outputs the results of CPC2 as a backup. In this case, CPC2 predicted 2573 putative novel ncRNAs.

If the user wishes to shorten the list of candidate ncRNAs to be further studied, several options exist. For example, a subsequent differential expression analysis that relates to physiological conditions that were compared can be performed. This facilitates deciding which novel transfrags most likely relate to a condition of interest and will be targets for further investigation. Another option is the addition of further restrictions, like transfrag length or number of exons, which would rather exclude lncRNAs. This will also reduce the number of interesting candidate transcripts.

DATA AVAILABILITY

Pinc will be accessible as a Nextflow pipeline on GitHub if the necessary programs are already installed (https://github.com/brummetheus/pinc) or as a Docker image to be able to run it independently on the user's preferred operating system.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

FUNDING

Christian Doppler Research Association; Austrian Federal Ministry for Digital and Economic Affairs; National Foundation for Research, Technology and Development; Novozymes. Funding for open access charge: TU Wien.

Conflict of interest statement. The *T. reesei* strain Iogen-M10 is a proprietary strain of Novozymes. C. Girod and D. Yaver are former employees of Novozymes.

REFERENCES

1. Christov,C.P., Gardiner,T.J., Szüts,D. and Krude,T. (2006) Functional requirement of noncoding y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol.*, **26**, 6993–7004.
2. Statello,L., Guo,C.-J., Chen,L.-L. and Huarte,M. (2021) Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.*, **22**, 96–118.
3. Sun,Z., Nair,A., Chen,X., Prodduturi,N., Wang,J. and Kocher,J.-P. (2017) UClncR: ultrafast and comprehensive long non-coding RNA detection from RNA-seq. *Sci. Rep.*, **7**, 14196.
4. Zhao,Q., Sun,Y., Wang,D., Zhang,H., Yu,K., Zheng,J. and Zuo,Z. (2018) LncPipe: a Nextflow-based pipeline for identification and analysis of long non-coding RNAs from RNA-Seq data. *J. Genet. Genomics*, **45**, 399–401.
5. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
6. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
7. Kovaka,S., Zimin,A.V., Pertea,G.M., Razaghi,R., Salzberg,S.L. and Pertea,M. (2019) Transcriptome assembly from long-read RNA-seq alignments with stringtie2. *Genome Biol.*, **20**, 278.
8. Pertea,G. and Pertea,M. (2020) GFF utilities: gffread and gffcompare. *F1000Research*, **9**, 304–304.
9. Kang,Y.-J., Yang,D.-C., Kong,L., Hou,M., Meng,Y.-Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
10. Wang,L., Park,H.J., Dasari,S., Wang,S., Kocher,J.-P. and Li,W. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
11. Youden,W.J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.
12. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
13. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
14. Martzy,R., Mello-de-Sousa,T.M., Mach,R.L., Yaver,D. and Mach-Aigner,A.R. (2021) The phenomenon of degeneration of industrial trichoderma reesei strains. *Biotechnol. Biofuels*, **14**, 193.
15. Margolles-Clark,E., Ihnen,M. and Penttilä,M. (1997) Expression patterns of ten hemicellulase genes of the filamentous fungus trichoderma reesei on various carbon sources. *J. Biotechnol.*, **57**, 167–179.
16. Castro-Piedras,I., Sharma,M., Brelsfoard,J., Vartak,D., Martinez,E.G., Rivera,C., Molehin,D., Bright,R.K., Fokar,M., Guindon,J. *et al.* (2021) Nuclear dishevelled targets gene regulatory regions and promotes tumor growth. *EMBO Rep.*, **22**, e50600.
17. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
18. Kalvari,I., Nawrocki,E.P., Ontiveros-Palacios,N., Argasinska,J., Lamkiewicz,K., Marz,M., Griffiths-Jones,S., Toffano-Nioche,C., Gautheret,D., Weinberg,Z. *et al.* (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.*, **49**, D192–D200.